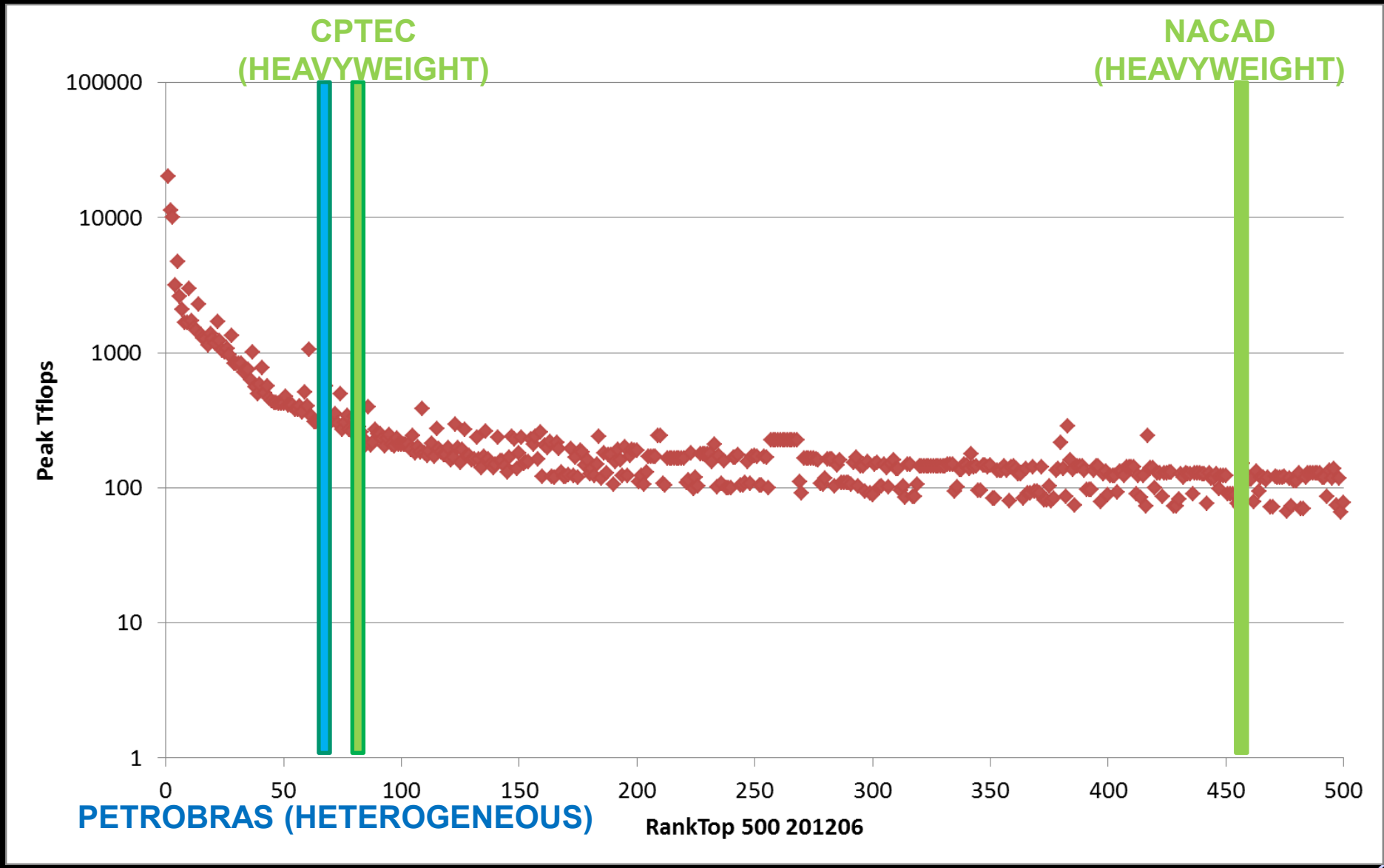
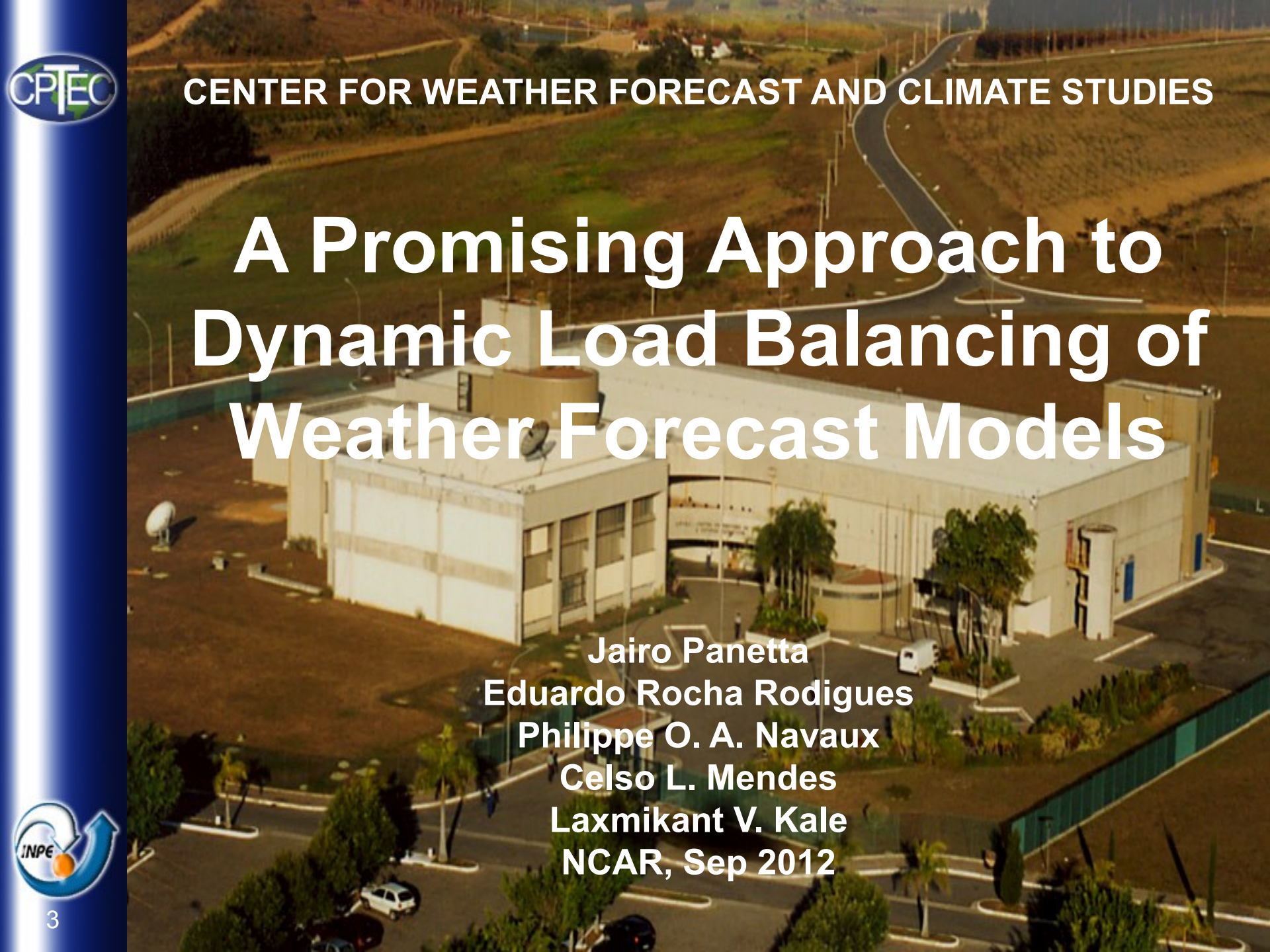
HPC @ Brazil

Jairo Panetta

ITA/IEC
Petrobras/E&P
INPE/CPTEC

Brazil at Top500 Jun 2012



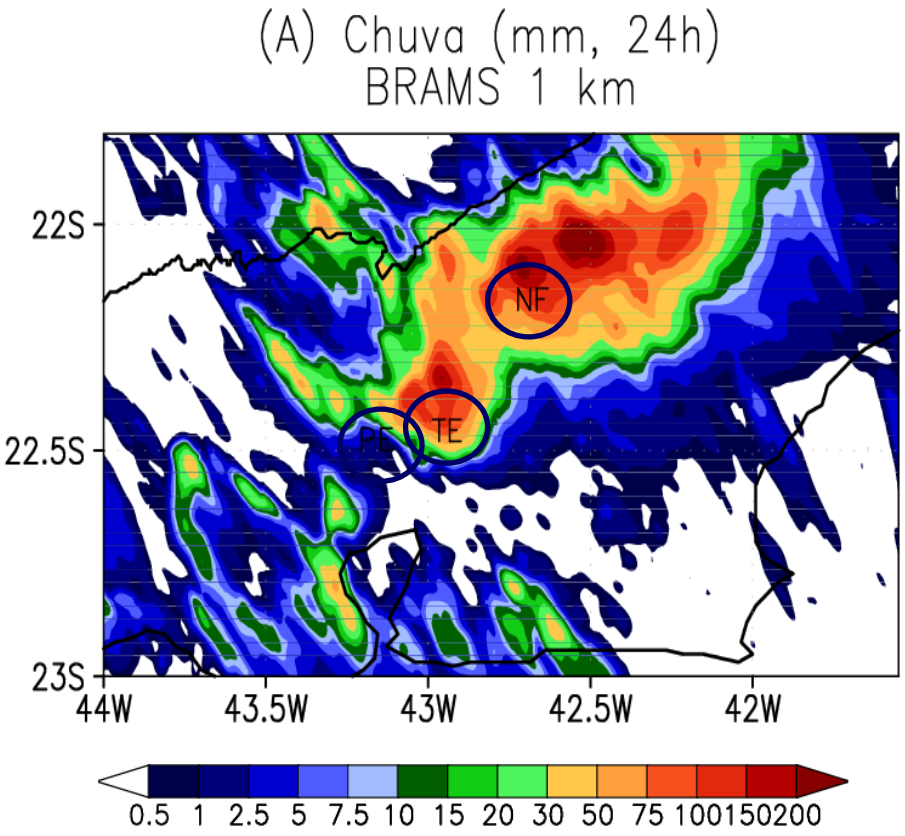


CENTER FOR WEATHER FORECAST AND CLIMATE STUDIES

A Promising Approach to Dynamic Load Balancing of Weather Forecast Models

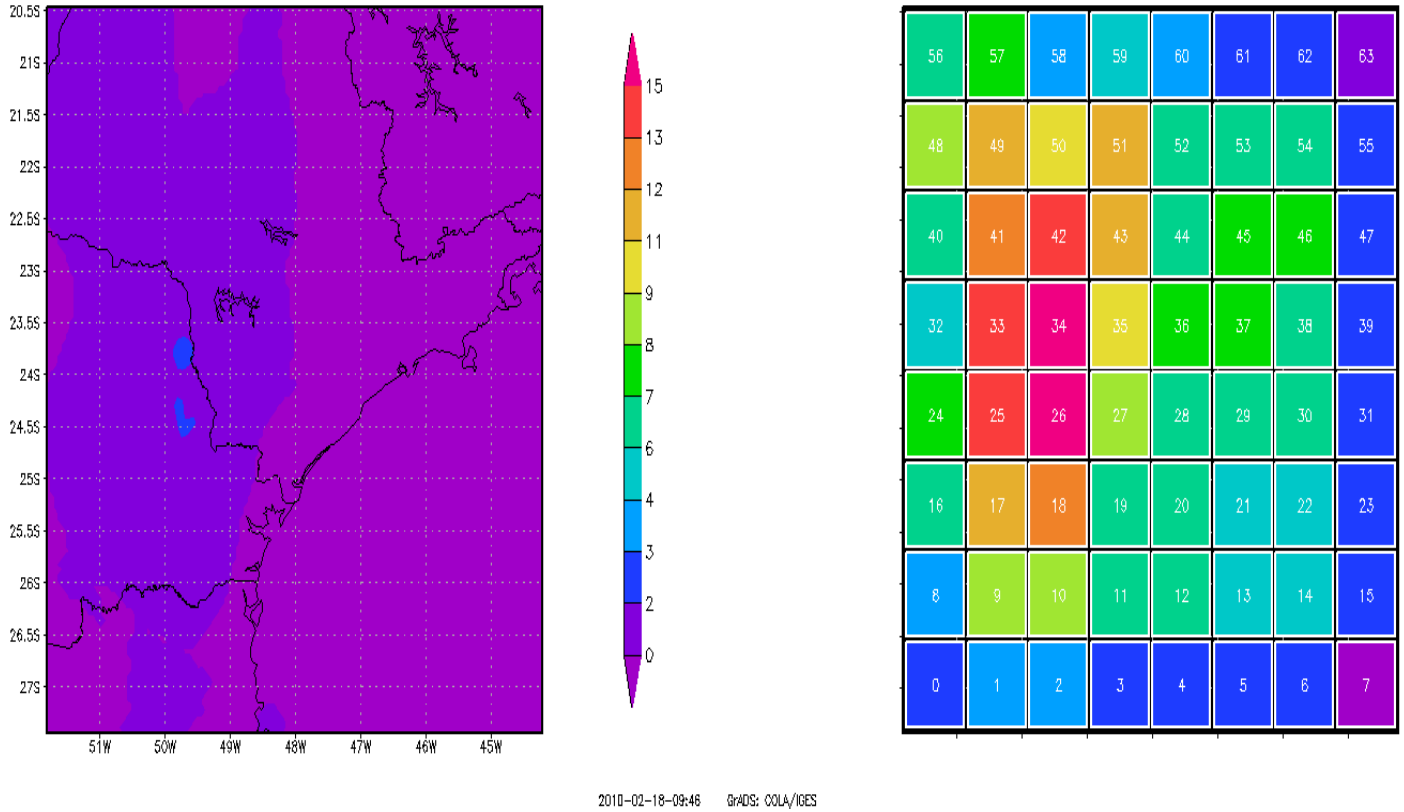
Jairo Panetta
Eduardo Rocha Rodigues
Philippe O. A. Navaux
Celso L. Mendes
Laxmikant V. Kale
NCAR, Sep 2012

BRAMS: Full Microphysics
(8 categories of water)



Accumulated Precipitation (24 hrs)		
City	Measured	Forecasted
Nova Friburgo (NF)	162	158
Teresopolis (TE)	78	88
Petropolis (PE)	7	4

The Problem



Dynamic Load Imbalance Limits BRAMS Scalability

brams.cptec.inpe.br

Desired Solution

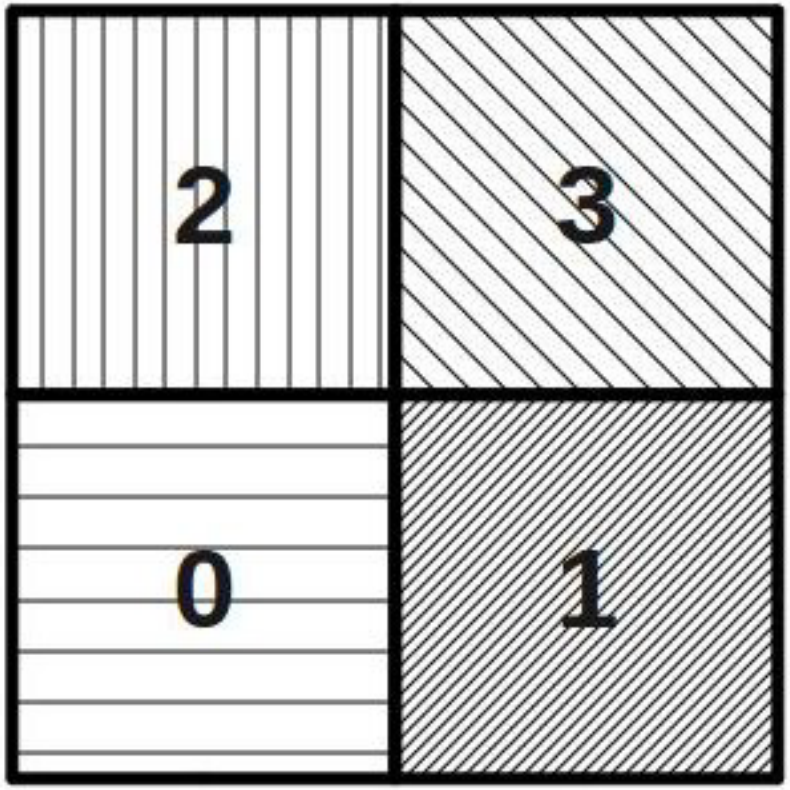
Automatic load balance with minimum (zero?) code intrusion

Research Strategy

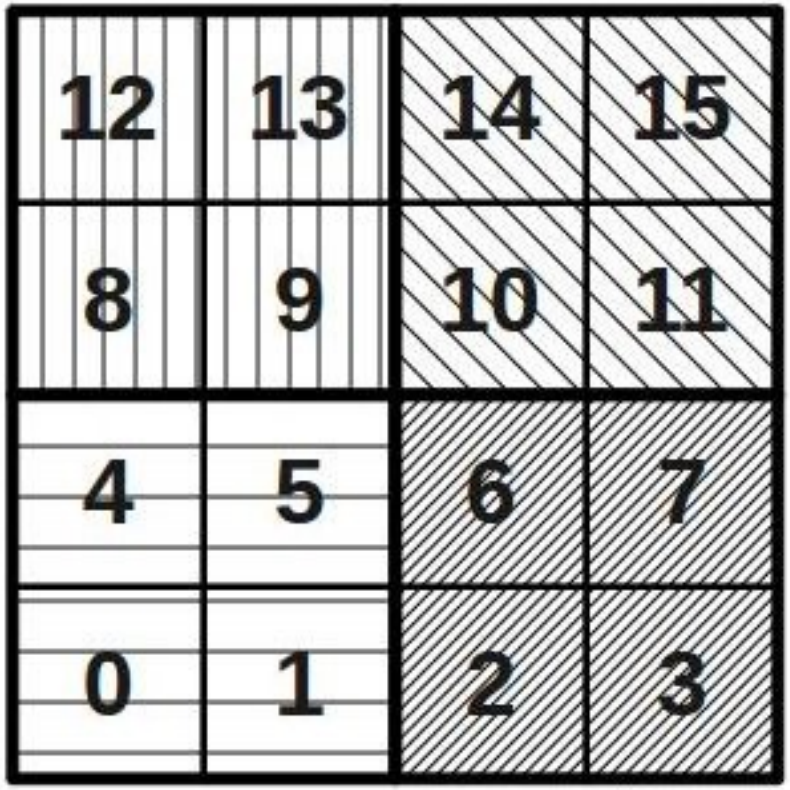
- Over-decompose the domain
  - More MPI ranks than real processors
  - Processor virtualization
- Move MPI ranks across real processors to balance the load
  - Use AMPI, an MPI library build on top of Charm++ (charm.cs.uiuc.edu)
- Explore:
  - Virtualization costs and benefits
  - Load Balancing Algorithms
  - Triggering factor to balance the load

Virtualization Costs and Benefits

Processor Virtualization



4 MPI ranks
4 real processors



16 MPI ranks
4 real processors

Virtualization through AMPI

- AMPI implements MPI ranks as user-level threads
- AMPI build-in scheduler keeps a thread executing until it is blocked (e.g., waiting for communication)
  - Overlaps communication and computation
- AMPI ranks are “migratable” and AMPI has a built-in set of thread migration algorithms
- But not all codes can use AMPI:
  - Source code cannot have static or global variables
  - Change gfortran and runtime libraries to generate (and run) code that supports Thread Local Storage at user-level threads

Results: Virtualization

- BRAMS at 10M Grid Cells, 1.6 km resolution, 4 hours of simulated time, 6 seconds time-step
- 64 real processors @ Kraken

Virtual Processors	Execution Time (s)	
1x64	4970	*no virtualization*
4x64	3857	
16x64	3713	*fastest with virtualization*
32x64	4437	

- By improving CPU utilization (overlapping computation with communication)
- By improving cache utilization (smaller sub-domains)

Load Balancing

Source Code Changes

- Only one line of BRAMS source code was modified (introduced), to invoke the load balancer:

if (<triggering factor>) call MPI_Migrate()

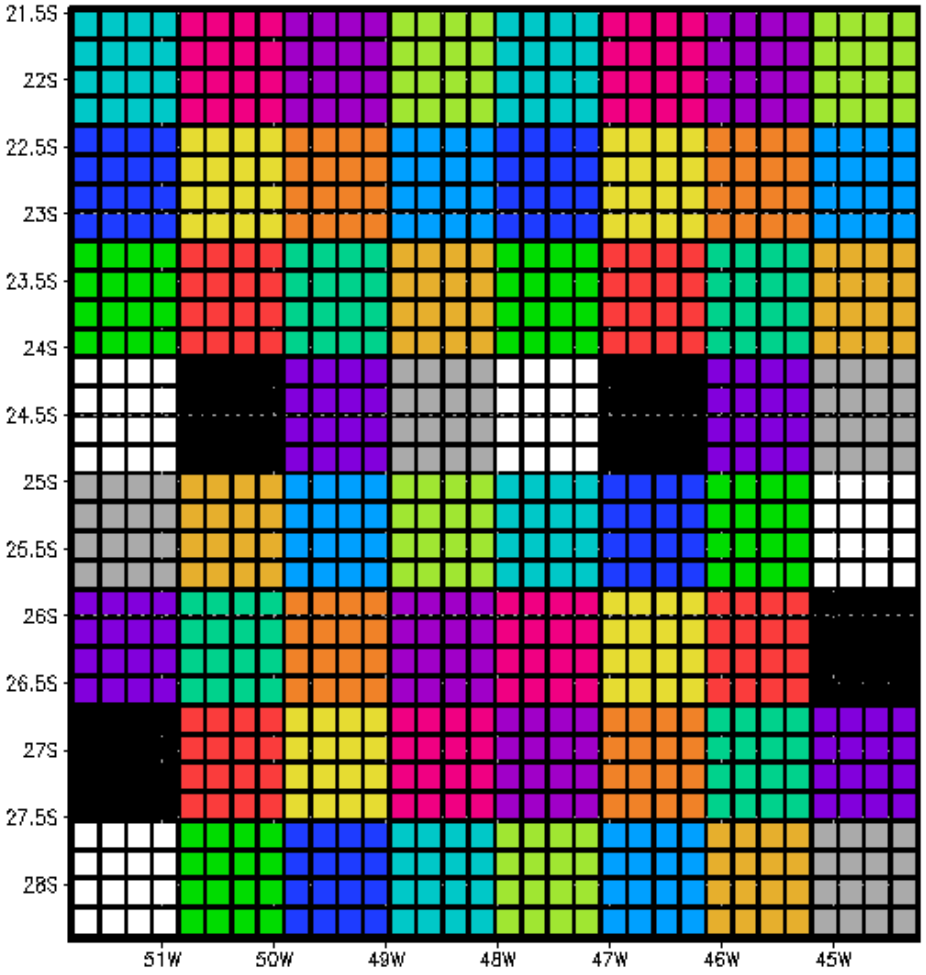
Load Balancing Algorithm

- An LB algorithm should achieve:
  - Fast execution
  - Good Load Balance after rebalancing
  - Low Communication Cost after rebalancing
- Key: Develop an algorithm that keep communicating ranks together after migration
  - Hilbert Space-filling curve
- Start with a fixed triggering policy
  - Once every simulated hour

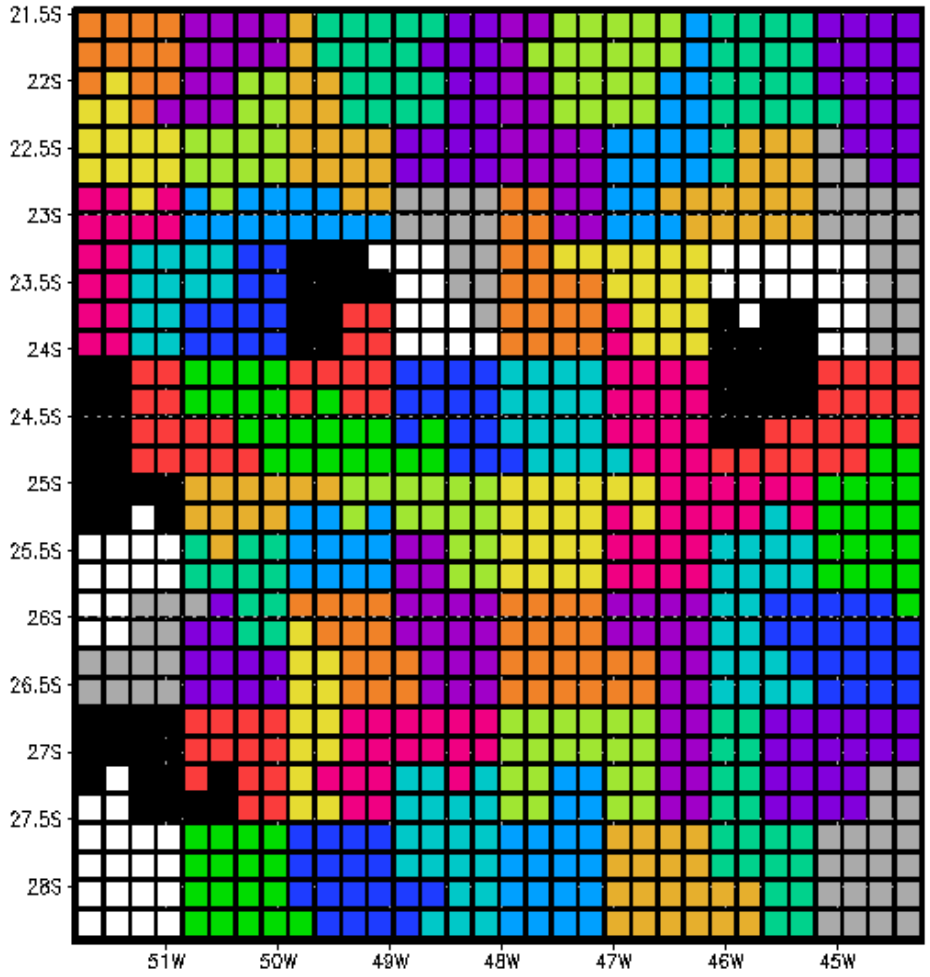
Results: LB Algorithm

LB Algorithm	Execution Time (s)
No OvrDec	4987
OvrDec, no LB	3713
OvrDec + Hilbert	3366

Initial Mapping of MPI ranks to 64 processors (8x8)



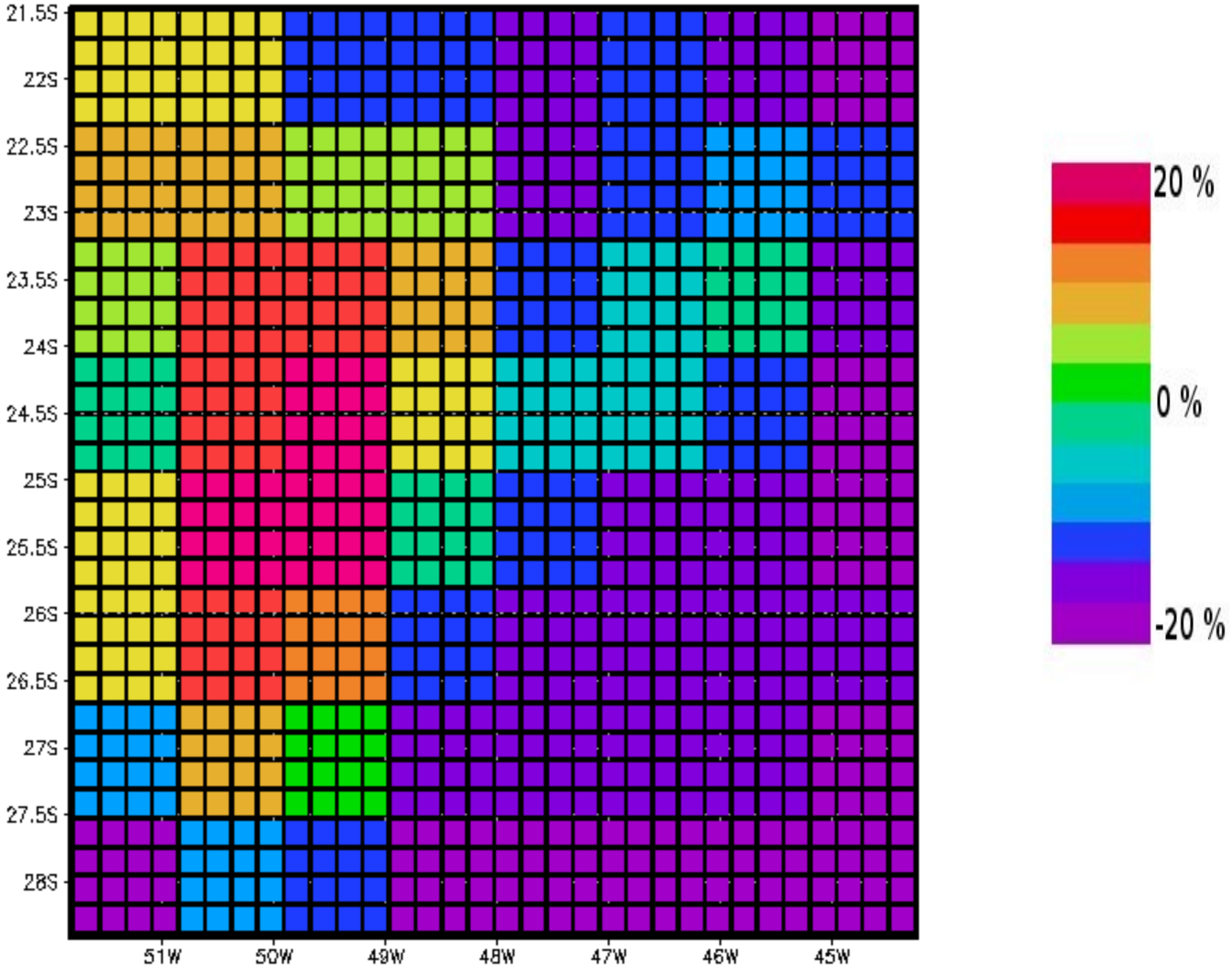
Mapping after rebalance at one hour



GrADS: COLA/IGES

2011-09-10-20:56

Load imbalance on the first hour prior to rebalance

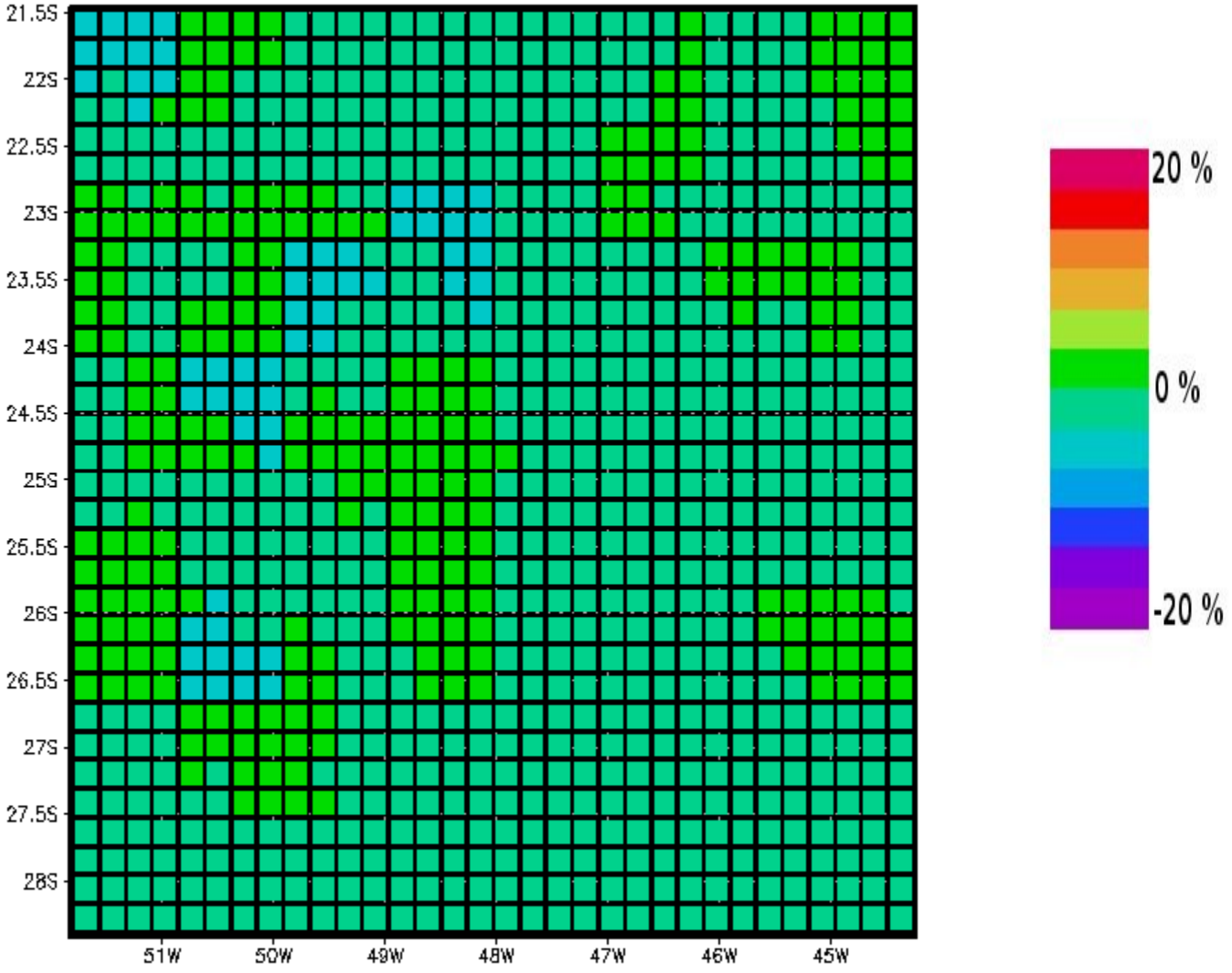


GrADS: COLA/IGES

2011-09-11-15:31

brams.cptec.inpe.br

Load imbalance on the first hour after rebalance



GrADS: COLA/IGES

2011-09-10-21:31

Triggering Policy

Triggering Policy

- Measuring load imbalance is cheap
- Migrating tasks is more expensive
- Policy: migrate whenever future execution time can be reduced (potentially)

Results Summary

Code Feature	Exec Time (s)
Original	4987
Include Over Decomposition	3713
Include Hilbert Load Bal.	3366
Include Automatic Trigger	3128

Conclusions

- There is evidence that processor virtualization and dynamic load balancing are beneficial
- Negligible source code changes
  - Provided no static or global variables
- Persistent over new parameterizations
- Further test cases are required...

Bibliography

- Rodrigues, E. R. et al, “A Comparative Analysis of Load Balancing Algorithms Applied to a Weather Forecast Model”, SBAC-PAD 2010
- Rodrigues, E. R. et al, “Optimizing an MPI Weather Forecasting Model via Processor Virtualization”, HiPC 2010
- Rodrigues, E. R. et al, “A New Technique for Data Privatization in User-level Threads and its Use in Parallel Applications”, SAC 2010
- Rodriges, E.R. “Dynamic Load Balancing: A New Strategy for Weather Forecast Models”, PhD Dissertation, PPGC, UFRGS 2011


THANK YOU